·数字史学研究·

编者按:数字人文的话题受到了学界持续的关注,甚至被评为2018年的十大学术热点之一,但数字人文研究的项目落地还面临各种考验。为此,《南京大学学报》编辑部与南京大学历史学院数字史学研究中心在2019年1月举办了一次"融合案例研究的数字史学再反思"小型讨论会,本次笔谈是此次研讨会的一个成果展示。邱伟云通过长期的学术实践,总结出数字史学与传统史学应该具有的良好互动关系。梁晨则以宏大的视角、详实的案例让我们领略到了数据库之于历史研究的强大助力。蒋勤以石仓文书为例,为我们显示了从无到有构建一个专业数据库的过程。王涛则再次呼吁中国的世界史学者拥抱数字史学,同时也分析了数字史学研究在中国的学术语境中遭遇到的难题。几位学者从多维角度展示了数字史学的无限可能与魅力。我们希望借此呈现数字史学这一新方法给学术研究带来的价值。

验证、修正、创新:数字史学方法的三重功能

邱伟云
(山东大学 历史文化学院,济南 250100)

"数字史学(Digital History)"与目前学界较常使用的"数字人文(Digital Humanities)"概念虽有着不同的指称,但基本上两个概念具有含摄关系,亦即数字史学包含在数字人文概念范畴中。那么,为何要特别标举数字史学概念?其主要作用是希望通过聚焦概念,以强化和推进数字技术在史学研究中的运用。王涛曾对数字史学进行如下定义:数字史学重视网络的利用与展示,着重数据库的建设,目前比较成熟的研究法是以GIS(地理信息系统)运用为主,另有计量分析与文本分析方法。① 从上述定义可掌握数字史学的大致意涵,亦即运用数字技术去推广或进行历史研究等相关工作。

虽然运用数字方法进行历史研究的源头可追溯到很早,但因某些因素,一直都没有成为史学研究法中固定的一环,近来则因为结合大数据时代来临的巨浪方能涌现,看来略有野火燎原之势,各领域都有不少学者进行数字史学相关研究,也有诸多示范性成果。在未来必将走向全数字化时代的体认下,学界也开始正视数字史学的发展,并思考这种新方法可能对传统史学研究带来的冲击与贡献。正是在上述脉络下,近几年在海内外都召开不少会议进行相关讨论。当我们从知识社会学角度去观察这些会议中不同领域学者对数字史学研究的有关意见时,可以发现,大多数师友都不曾受过系统的数字史学研究法训练,也非长期进行数字史学研究工作者,因此对数字史学的评议位置,多是从外围进行的接触与想象。这是一个有趣的现象,因为就这些从外围进行的观察意见来看,消极的不安与质疑意见是多于正面肯定意见的。然而这些从外围来的消极意见,事实上是很有意义的,因为这些意见有助于让数字史学工作者掌握目

① 王涛《"数字史学":现状、问题与展望》,《江海学刊》2017年第2期。

前学界对数字史学的看法，这对推进数字史学研究发展具有正面帮助。亦即只要能较好地回应这些问题，就能使数字史学方法为更多人所接受，更能稳健地发展。

笔者梳理了过去曾参加过的相关会议中与会师友对数字史学曾提出的意见，可归纳出一个常见、重要且必须回答的问题，那就是：数字史学在研究过程中，需研究者花大量时间清理资料数据/文本数据，之后还须学习数字技术以进行计算，而在复杂计算后还须从数据线索中找到问题意识，进而确认文本与解读文本，最后才能综合地给出研究结论；相较过去史学研究法，数字史学工作者所花费的时间与力气比过去还多，但研究结果却看似无法超越过去史学研究者的观点，至多只是验证过去的研究结论，那么，为何史学工作者还要花更多的力气去学习数字技术并进行数字史学研究呢？

上述这个问题很有趣也很重要，换个问法会使问题意识更明确，亦即：为何数字史学研究结果总是不能让传统史学工作者耳目一新？这里面的问题关键究竟在哪里？难道是数字史学方法真的没有太大用处？或是数字史学真的就只是一种只能重复传统史学研究结论的方法吗？

为了让数字史学研究法具有存在的合法性，数字史学工作者应肩负起指出方法优势与价值的重任，才能让学者们了解其作为一门新方法的合理性与有效性。那么，该如何标举出数字史学方法的价值呢？笔者认为，可以尝试在每个数字史学研究项目最后，都在结论部分特别标举出验证、修正与创新等三个面向，透过自觉的辨别，更为聚焦地思考与凸显数字史学方法的长处与优势，这是在数字史学理论发展之初，证明其作为新方法实具有合法性与有效性时的重要工作。以下笔者就从三个运用文本探勘技术进行史学研究的数字史学研究案例出发，尝试归纳数字史学方法的三重功能：验证、修正、创新，并从实际研究案例中进行观察与反思，希望能借此建立数字史学作为一种新史学研究方法的主体性，夯实数字史学发展的基础。

谈及数字史学研究法的第一种功能应是“验证”，即透过数字技术，从巨量数据中，借由计算与分析，最后量化地验证过去史学研究者曾提出过的研究结论。例如在中国近代主义概念研究方面，胡适曾在 1919 年提出“多谈些问题，少谈些主义”的论点，基于此，著名历史学家王汎森先生运用思想史方法，探究中国近代“主义”概念的发展，透过为数不少的重要思想史材料，指出“五四”后有“主义”风靡知识界和大量的“主义化”现象①。对于这个问题，数字史学工作者詹筌亦与王乃昕曾尝试运用数字技术，对中国近代主义概念进行数字史学研究②。他们从计算思维出发，提出具有数字史学特征的问题意识：当时主义风靡的程度能否用量化方式证明？有可能找到中国近代产生多少种主义吗？两人正是基于上述量化的问题意识，运用数字技术进行巨观考掘，以包含一亿两千万字的“中国近现代思想史专业数据库（1830－1930）”为研究对象，经过计算，结果发现 1896－1928 年间，在数据库巨量文献中共有 1 680 种主义，并且还给出每年主义种类的使用数量、前 20 种重要主义的排序，从量化角度勾勒出“主义”概念的历时性发展轨迹，量化地证明了近代中国“主义化”的风靡程度。此一研究的人文意义在于透过语言词汇的量化证据，证明了“主义”确实风靡中国近代知识界，而非仅是历史学者个人主观对历史材料的长期阅读印象所致，此一研究即体现了数字史学研究法所具备的“验证”功能。

① 王汎森《“主义时代”的来临——中国近代思想史的一个关键发展》，《东亚观念史集刊》第 4 期，台北：政大出版社，2013 年。

② 项洁等主编《数位人文在历史学研究的应用》，台北：台湾大学出版中心，2011 年，第 219－245 页。

数字史学研究法的第二种功能是“修正”，即透过巨量资料的计算分析，对过去史学研究结论进行补充修正工作。如在《新青年》的思想转型研究方面，过去一般认为，陈独秀在创刊号上就发表了《敬告青年》一文，文中指出中国富强的关键系于青年，大部分学人就此认定《新青年》创刊目的就是教育青年，以青年为主。对于“青年”概念是否一开始就为《新青年》主要概念这个问题，由金观涛、梁颖谊、姚育松、刘昭麟等 4 位具有历史、统计、计算机等学科背景的学者合作，他们利用 PAT – tree 数字撷词技术及根据 Zipf – Mandelbrot 模型计算出的理论曲线，统计出《新青年》11 卷中各卷的“关键词”①。该文透过巨观且复杂的数字计算后指出 “青年”概念并非《新青年》杂志第 1 卷中的关键词，要到第 2 卷才是，第 1 卷中的关键词是“国家”与“政府”，这就促使研究者去思考分析为何会出现此种不同于过去史学研究结论的现象，进而产生新的问题意识。该文进行史料分析后指出，《新青年》之所以并未从第 1 卷开始就重视“青年”概念，是因为当时国家主义尚未幻灭，因此“国家”与“政府”概念还凌驾于“青年”概念之上，要到第一次世界大战后的第 2 卷开始，知识分子才从国家主义中醒悟，真正走向关注青年、女性等个人自觉的道路，这时“青年”一词才会成为关键词。上述这一结合统计与计算机方法的数字史学研究，修正了《新青年》杂志一开始就是以“青年”为主的说法，并且还补充指出了“青年”概念能在第 2 卷后涌现，实与欧战有密不可分的关系，此一研究即体现出数字史学研究法所具备的“修正”功能。

最后，数字史学研究法的第三种功能是“创新”，前述已指出数字史学视野的优势是能进行巨观且复杂的计算，这是传统史学视野所无法触及的面向，因此要呈现数字史学的创新功能，可从巨观与复杂计算视野出发，提出新的研究问题。这样的创新案例，可用笔者正在进行的“中国近代‘人’观的百年嬗变研究”为例。虽然中国近代“人”的观念发展很重要，但由于近代谈及“人”的史料异常巨量与复杂，因此一般研究者实难以驾驭这一问题，致使过去研究多从“个人”概念出发，以小见大，以“个人”概念作为“人”观的代表。对中国近代“人”的观念发展进行讨论，指出中国近代有一从“大我”走向“小我”的发展轨迹，这是在传统史学研究法下能够提出与处理的问题与解答。然而，“个人”概念并不能代表整体“人”观，因此目前对于“中国近代‘人’观的百年嬗变”此一问题还没有能取得较为系统与全面性的处理。如今，凭借数字史学技术能处理巨观且复杂问题的优势，数字史学工作者得以尝试提问与解答过去难以研究的巨观历史问题了。

针对上述命题，在数字史学视野下，研究者从计算角度进行问题转译后的量化问题意识是：中国近代百年之间以“人”为前缀词（如人类）的词汇有哪些？为后词缀（如白人）的词汇又有哪些？又以哪些“人”的词汇用得最多？找到上述这些问题的答案，就能揭示中国近代思想转型时代中“人”的多元观念系统结构。从上述问题意识中可以很快发现，中国近代“人”观的发展无法使用传统的思想史研究法去回答，因为即使记忆过人，半年可以读完一亿字的近代文献，但要过目不忘并记下所有“人”的词组还是不太可能的，然而若是使用数字技术，就能很快地完成这个工作。因此，如上述指出的巨观且复杂的问题，正是适合与专属于运用数字史学新方法去处理的新的史学研究问题。

综上，笔者以前文曾提到的“中国近现代思想史专业数据库（1830 – 1930）”作为研究对象，以数字技术计算出一亿两千万字的近代重要政治思想史文献中，所有以“人”作前后缀的

① 金观涛、梁颖谊、姚育松、刘昭麟《统计偏离值分析于人文研究上的应用——以〈新青年〉为例》，《东亚观念史集刊》第 6 期，台北：政大出版社，2014 年，第 327 – 366 页。

词汇，并将这些词汇的词频计算出来后进行排序，可勾勒出数据库中的“人”观的系统结构。在前缀词的部分依词频高低顺序为：人民、人人、人类、人心、人生、人口、人才、人数、人物、人员等，词频超过 500 次以上者有 39 个词；在后词缀的部分依词频高低顺序为：国人、人人、英人、工人、俄人、华人、西人、个人、外人、法人等，词频在 500 次以上者共有 106 个词。从计算后得到的前、后缀词的词频列表中，可以很快全面掌握数据库中所收的百年报刊文献里提及的“人”观整体系统结构，而这些数据线索可打开研究者的思路与问题意识，如在前缀词中以“人民”为最高频使用词汇，后缀词中则是以“国人”一词使用最多，更胜于“个人”，这些结构性数据信息即可刺激研究者兴发问题意识，进而进行更多的史学探索。

由于前面的词频计算不具时间序列，无法观察到“人”观的发展轨迹，因此凭借数字史学方法的计算优势，接着可将上述百来个“人”的词汇，在 1830 – 1930 年间于数据库的使用情况全面地描绘出来。这与过去在检索思维下只透过查询一个、两个乃至 10 个概念（词）的发展轨迹并进行比较已有所不同。

在“数据驱动（Data Driven）”思维下的数字史学研究法中，是整体概括所有“人的词汇”并进行计算，不同于过去是由研究者主观选择重要概念进行讨论，容易受到研究者主观视域的选择性限制。透过 CUSUM 统计方法，可以立即描绘出百来个“人的词汇”的百年使用比例变化轨迹，并从发展趋势的近似性特征中，进一步进行群聚计算，即可在未加入研究者主观意识下，快速将百来个“人的词汇”进行时间序列分群。从整体数据结构中的数据线索出发，得以兴发许多不曾想过的问题意识，进一步可再对相关史料进行阅读分析，完成数字史学的研究工作。

笔者在计算数据线索中，即发现几个有趣并值得继续探问的思想史问题，如数据显示自晚清到民初有一从“人人→人民/人群→人类”的发展，这代表的是“人”观的去国族化与世界主义化；又或者有一从“人心/人欲/人情→人性→人格”的发展，这即是受到“人”观的泰西伦理化影响所致；又或有一从“夷人→洋人/西洋人→西人→欧洲人/欧人/外国人→欧美人”的发展，显示的是“人”观的万国化。上述这些数据线索是透过巨量复杂的数据计算下所得出的巨观结构，数字史学工作者的任务，就是从这些数据线索一一出发，进一步确认并探究上述数据线索的正确性及其带有的历史意涵，最后才能从总体上揭示中国近代“人”观的百年嬗变轨迹。

从上述创新研究案例中可见，数字史学方法存在的合法性，是建立在与过去研究方法不同的巨观与复杂计算之上，不断地寻找与提出只有运用数字技术才能提出的史学问题并给出答案，便是回应人们对数字史学方法合法性质疑的最好办法。

最后想指出，数字史学与传统史学研究乃是相辅相成的关系，可相互对话与补充，因此学界或可采用更宽容的态度去理解与接受数字史学方法。作为大数据时代中的一种史学新范式，数字史学方法确有其存在的必要，期盼未来有志从事数字史学工作者，能积极地提出更多专属数字史学能提出的好问题，为数字史学研究提供更多经典的成功案例，进而共同推进数字史学的研究发展。